ASp
la revue du GERAS

39-40 | 2003
Varia

From meaning to words and back: Corpus linguistics and specialised lexicography

Geoffrey Williams

Electronic version
URL: http://journals.openedition.org/asp/1320
DOI: 10.4000/asp.1320
ISBN: 978-2-8218-0392-3
ISSN: 2108-6354

Publisher
Groupe d'étude et de recherche en anglais de spécialité

Printed version
Date of publication: 1 May 2003
Number of pages: 91-106
ISSN: 1246-8185

Electronic reference
Geoffrey Williams, « From meaning to words and back: Corpus linguistics and specialised lexicography », *ASp* [Online], 39-40 | 2003, Online since 11 May 2010, connection on 01 May 2019. URL : http://journals.openedition.org/asp/1320 ; DOI : 10.4000/asp.1320

From meaning to words and back: Corpus linguistics and specialised lexicography

Geoffrey Williams

Introduction

1 Words do not have meanings, meanings have words. This may seem obvious, and is the basis of the Saussurean notion of the *arbitraire du signe*, but it is often far from our everyday attitude to language. The contextualist school of thought that derives from Firth (1890-1960) puts flesh onto the notion of arbitrariness in declaring that the meaning of a word can only be fully appreciated in context, the context is primordial. This poses a major problem in dictionary writing as an entry is always out of context. Meaning thus represents a challenge to both the lexicographer and the dictionary user. For the lexicographer meaning must be transferred from context to the dictionary entry using a metalanguage that is sufficiently clear to the user. For the user, the challenge is to transfer meaning from the dictionary to the text, and in writing from the dictionary to a new context.

2 A revolution in dictionary making came with the development of corpus linguistics, built on the contextualist view of meaning, and its transfer to lexicographical practice through the COBUILD dictionaries. Corpus linguistics meant analysis of words in context to demonstrate use in context, which entailed changing the dictionary format so as to enable the transfer of this contextual knowledge back to the user. This has created a revolution in both mono~ and bilingual dictionaries. The contextual approach is now transforming even terminology as, in such real life usage, conceptual rigidity no longer holds.

3 The aim of this paper is to trace the changes in dictionary design that corpus linguistics has brought about and to show how approaches initially developed for general language

reference dictionaries must be adapted to specialised usage if we are to help users transfer their meanings into words on the page.

1. The Dictionary: A tool with many faces

4 Dictionaries come in many forms, and serve a wide variety of purposes in addition to that of teaching. "*Dictionary*" is indeed a polysemous word covering works as different as historical dictionaries, such as the Oxford English Dictionary (OED), and highly encyclopaedic works as the Oxford Dictionary of Biochemistry and Molecular Biology (ODBMB). The first is a classic language dictionary, the second is more encyclopædic in nature and concerned with terminology; what they have in common is a tendency to present words as discrete items in alphabetical order. This semasiological presentation may not be the best, but it is what we have come to expect of a dictionary. The only exceptions to this alphabetical rule in our daily usage tend to be the onomasiologically organised thesauri, such as Roget's thesaurus.

5 The wide variety of dictionary types means that it is far from easy to define the concept precisely, although we all know what we "mean" by dictionary and can recognise one when we see one. According to the Oxford Advanced Learners Dictionary (OALD), a dictionary is:

> (**a**) a book that gives the words of a language in alphabetical order and explains their meaning or translates them into another language. (**b**) a similar book that explains the terms of a particular subject. (OALD)

6 In fact the OALD gives us three definitions. The OED and OALD both fall into the "a" category as they seek to explain the meanings of words, whilst the ODBMD is clearly a member of the "b" category. The "a" category is, however, divided into two as we have both monolingual dictionaries, as in my examples, and bi- or multilingual dictionaries, such as the Roberts and Collins Senior (RCS).

7 The extremely broad definition of the OALD leaves much unsaid, but as we all "know" what a dictionary looks like, this is generally not considered a problem. Unfortunately, however, in both general and metalexicographical terms, there can be hidden problems. General users are invariably unaware as to how dictionaries are put together and what audiences they address. They also tend to have a "fixist" attitude to the meaning of words, which causes them to forget or ignore the evolution of language. These factors can lead our students to rely on dictionaries they find at home, whether mono- or bilingual, which are hopelessly out of date, and which lead them to use general language meanings in specialised contexts, often with highly amusing results. These problems often arise from a lack of knowledge of dictionary types and their uses, a factor which is widespread even among language teachers. Obviously knowledge of dictionary typology is crucial for the lexicographer; an detailed analysis of the problem from the viewpoint of a metalexicographer can be found in Béjoint (1994 2000), required reading for anyone interested in lexicographer or language teaching. We do not need to go into detail here.

8 The difference between mono- and bilingual dictionaries is obvious and although the pitfalls of using the latter are well known in language teaching, we shall not consider them here. It must however be pointed out that the bilingual dictionary, like its monolingual counterpart, has also gone through revolutionary change with the advent of the computer. Sue Atkins, a prime mover in the modernising of bilingual dictionaries has

discussed this in detail (Atkins 2002). Here we are concerned with monolingual dictionaries, both general and specialised.

9 If we take the definition of a monolingual dictionary given in the OALD, we miss an important point. The father of English Dictionaries, the OED, is in a line that can be traced back to Johnson's 1755 Dictionary. These are dictionaries which, whilst not being fully prescriptive, are normative. Most importantly, the OED has developed into a historical dictionary concerned with showing the evolution of the English language, largely through literary usage. Its aims go well beyond simply giving the meanings of words. The smaller COD on the other hand is a reference work, but like most shorter works is heavily reliant on the larger dictionary, in this case the OED. Similarly, if we looked up the definition of the word "dictionary" in the COD and compare it with that of the OALD, we would find the two are almost identical. Again, the OALD relies on its big brother; it is simply far too time-consuming and expensive to start a dictionary from zero. Yet, although the definitions may be similar, the COD and OALD are fundamentally different. The COD addresses native speakers who generally consult dictionaries to check spellings or to check the meanings of words about which they are unsure, it is thus designed essentially for passive usage. The learner's dictionaries, such as the OALD, were initially designed for non-native speakers and sought to combine two roles, decoding and encoding. The decoding aspect is the passive one where learners look up ill-understood words so as to elucidate their meaning in context. The encoding aspect goes well beyond simply spelling and must enable the writer to produce understandable language, which entails giving examples of usage and carefully encoded grammatical information.

10 Contrary to general language dictionaries, works dealing with specialised usage are generally terminologically based and come in two forms, multi-lingual terminologies that mostly address the translator, and which are not the subject of this study, and the monolingual encyclopaedic dictionaries that address subject specialists and which present the essential background for a discussion of a possible specialised learner's dictionary. However, the approach of such dictionaries is entirely different to that of the learner's dictionary. The classic definition system used in most general languages works seeks to give an intentional analysis of a word, that is to say it gives the distinctive features of a concept whereas encyclopædic entries go well beyond this, giving detailed extensional information. The aim of dictionaries such as ODBMB is to fix and explain terms, they address only users of native speaker competence and make no attempt to show or explain usage. Such dictionaries are by their very essence prescriptive and do not set out to teach how to use these words in real-life situations.

2. The dictionary as a teaching tool

11 The advent of modern computing has revolutionised dictionary making by providing not only data-bases to allow more efficient handling of data, but also access to new forms of data and, for learner's dictionaries in particular, new ways of presenting that data. The revolutionary changes in these encoding dictionaries provide the main thrust of this study, but before looking at the nature of the revolution we must consider the relationship between teaching and lexicography.

12 Teaching and dictionaries have always been inextricably linked, from the English-French dictionary of Holyband of 1593, through to Hornby's *Oxford Advanced Learner's Dictionary,* first published in 1948. The story of the rise of ELT is told by Howatt (1984) and is one in

which we see advances in pedagogical practice, especially from the mid-war years, going side by side with advances in dictionary making. The rise of the ELT dictionary went hand in hand with the selection and explaining of the essential language needed by the second language learner.

13 The need for a controlled vocabulary for readers had become apparent in the 1930s and led to the work of West and his famous *General Service List of English Words*, published in 1953 (Howatt 1984). Hornby's work went beyond lists and the needs of lower-level learners to build a fully fledged dictionary for advanced learners. Hornby's dictionary was not simply a down-sized version of larger work, but a purpose built one compiled with the needs of the non-native speaker in mind. The OALD was to contain more that just individual words, but idioms and collocations, the fruit of Hornby's long collaboration with Palmer in Japan (Cowie 1998). With the OALD we have a dictionary turned to the needs of language production, with explanations and examples of word patterning. Later editions have benefited from the grammatical information brought to light in Quirk *et al*'s corpus-based *Grammar of Contemporary English* (Quirk *et al*.1985).

14 Hornby's dictionary set off the process which has led to there being a plethora of learner's dictionaries on the market place, dictionaries valued as much by advanced learners as by native speakers. In a recent study of dictionary use by second language users in the USA, McCreary and Dolezal (1999) found that use of learner's dictionaries lead to far better results than standard American college dictionaries, and that even the native speaker American control group would have benefited from similar works in avoiding the standard pitfalls of dictionary use. McCreary (2002) then went further with an in-depth study on American university students which showed that the students used poor dictionary use strategies, with poor results on difficult vocabulary. Given standard college dictionaries and learner's dictionaries, users of the latter were found to notably outperform users of the former. The learner's dictionary should by no means be seen as purely for non-native users.

15 In these learner's dictionaries, the presentation of word senses is unashamedly contextualist, words only acquire meanings in context, and therefore the dictionary must endeavour to show those contexts by showing real usage. The remaining problem was as to what words to include in the dictionary. Word lists, no matter how good, are subjective. Hornby had used the COD as the basis of the OALD, eliminating words he considered not useful for non-native students. This choice was based on his intuition and tremendous experience as language teacher, researcher and lexicographer. Whilst one should not underestimate the knowledge of a trained lexicographer, the resulting choices, both of words and the ordering of senses, is inevitably subjective. This obstacle could be overcome with the use of computers and electronic corpora.

3. Dictionary Making

16 Computers have had a major effect on dictionary making, the card files of traditional lexicography have disappeared into data bases which allow for easy stocking, transfer of data and above all cross-referencing. The use of SGML-based storage models has made the reformating of existing materials easy, simply a matter of changing the style sheet. The advent of the internet and cd-roms has meant new formats being developed with rapid user access to the data. Despite the criticism that many electronic dictionaries are only paper ones in an electronic format, online and cd-rom dictionaries do offer many

practical advantages over their paper counterparts. In turn computing has been helped by lexicography in that researchers in natural language processing (NLP) have had access to electronic material for analysis (Fontenelle 2002). This exchange will in turn be beneficial for both human and machine applications.

17 This computer technology has been adopted by all major lexicographical projects, but in certain areas an even bigger revolution has taken place in the nature of the source data. Corpus linguistics has transformed much lexicographical practice by providing access to vast amounts of authentic data. The first to realise this potential was John Sinclair and the COBUILD team at the University of Birmingham.

4. The COBUILD revolution

18 As we have seen, up until recently dictionary making has relied on the intuition of trained lexicographers in the analysis of material and the writing of definitions. Headwords were chosen on the basis of perceived importance and polysemy was treated in the same way. Even learner's dictionaries had grown over time with no clear criteria for the inclusion or exclusion of words. The COBUILD solution was bold; the team would build a dictionary from scratch based not on file cards, but on a large electronic corpus. It should be noted that from now on the word *corpus*, plural *corpora*, will only be used to refer to large electronic corpora assembled from very large quantities of authentic text

19 The COBUILD story has been related in detail in *Looking Up* (Sinclair 1987) the main points that consider us here are the criteria of headword choice, the analysis of meaning and the presentation of the dictionary entries.

20 The first revolution was the building of a reference corpus. Corpus in the sense used in corpus linguistics is a large collection of authentic texts that have been selected and organised following precise linguistic criteria (Sinclair 1996). The criteria for the development of reference corpora are now well established (Atkins *et al* 1992, Biber 1993.) with more and more becoming available in the world's major languages. In Britain, the original *COBUILD* corpus has grown into the monumental *Bank of English*, and is still expanding. The *British National Corpus*, built by a consortium of dictionary publishers, has established itself as a reference in corpus studies as an entirely annotated corpus. The first revolution was thus the corpus, next comes it exploitation.

21 Headword inclusion is a major problem for dictionary writers: what to include, what to exclude. The question is discussed with boring regularity on French television, whenever a new dictionary comes out, with fervent discussion as to what slang or buzz words have "entered" the language. In reality, their "entry" into a dictionary is often largely a subjective issue; the COBUILD team did not want subjectivity, but a reflection of reality. The result was that word selection would be based on corpus frequency, which means that usage of the word can be monitored over time.

22 The next stage was a rewriting of definitions based on corpus evidence rather than on previous dictionaries. This analysis not only held for lexical words, but also for "empty", grammatical words. Describing the determiner "the" as an easy word is clearly nonsense, it was necessary to show how it was used in context. Once the definitions written with polysemic words have been divided into "senses", the individual entries must be ordered. The decision here was not to order them by part of speech category or by intuitive notions of centrality of sense, but by frequency of use. This form of ordering is based on

the notion that sense and syntax are intimately related, a notion which has led Sinclair to declare that we must go beyond lexico-grammar to lexical grammar, a situation where the two levels become one (Sinclair 2002). Ordering by sense leads to problems of presentation, overcome in the COBUILD dictionaries by a side column giving grammatical and semantic information associated with each "sense".

23 Nobody claims that the COBUILD system is perfect, each dictionary publisher has its own house style, and dictionaries appeal very much to personal preferences of individual users, but the revolution is a fact of lexicographic life and no learner's dictionary can ignore the COBUILD revolution. Unfortunately, whilst the advent of better learner's dictionaries has had its effect on students studying English as a language, monolingual dictionaries have still not really penetrated the world of ESP.

5. Specialised dictionaries for ESP/EAP

24 Better dictionaries have not solved all our problems. The problem with many ESP students, particularly those in the sciences, is that they are not ready purchasers of dictionaries; at best they will use a bilingual dictionary and fall into all the false friend traps that are presented. Given that McCreary (2002) has shown that native speakers have poor dictionary skills, it can be assumed that non-native speakers will be no better off, especially if the examples given are not from their area of expertise as this requires a transfer of sense from general usage to a specialised context. At this point it might be interesting to see the sort of problems that McCreary found. Four main erroneous strategies were confirmed:

- *The 'Kidrule' Strategy*. - "the students assume that the tested word, the entry word, is semantically equivalent to one of the easier words in the definition". (*op.cit*: 194)
- *The 'choose the first definition' strategy*.
- *The 'superficial cognate' rule* (the malapropism creation strategy). – "When confronted with a 'hard word', think of a more familiar word (that may not necessarily be in the entry) that at first glance appears to be similar to the test word...". (*op.cit*: 196)
- *Choose the sexiest sense* – "...hop over the boring entries, and try to insert the sexy sense into the test word used in a sentence". (*op.cit*: 199)

25 All these strategies have been noted by other researchers, the second being the most prevalent. The source of the problem is obvious; no matter how hard we try a dictionary can only present a decontextualised meaning, it is the user who must operate a transfer of meaning from a dictionary source to a real text. It follows that the farther we are from the context that the user wishes to encode or decode the greater the risk of misunderstanding. Two solutions may be proposed; teaching better dictionary skills and preparing more contextually relevant dictionaries.

26 Improving dictionary skills is the strategy adopted by Campoy Cubillo (2002). Working with chemistry students, she got them to write their own dictionaries. This served two purposes; it enabled the students to familiarise themselves with dictionaries and it enabled the researcher to better understand how learners use dictionaries. Using classroom concordancing it would be possible to teach students to build small personal dictionaries by getting to grips with real language. However, although this may be useful on a small scale and would have pedagogical value, ESP students are not lexicographers and do not possess either the time or skills to build any but a basic lexis. Many students,

including those studying language sciences, have great difficulty in adapting to concordance analysis, which is after all a very particular skill. Obviously, in terms of real dictionary building, the ball remains in the camp of the lexicographer.

27 Norman (2002) has looked at existing specialised dictionaries and found that they are rarely corpus-based, tend to be written by field specialists who have no lexicographical training, and are resolutely prescriptive. These works are encyclopaedic in nature and are purely terminological in content. As Norman points out, the problems come frequently from semi-technical words that are often highly polysemic, his solution is a request for greater transparency in prescription and greater use of corpora in defining meanings. Whilst this might help with decoding, for most NN writers the major problems is one of encoding.

28 In Williams (2002) I pointed out that we cannot simply rely on lexicographers changing their working methods, all the more so as most specialised dictionaries are written by non-linguists with prescription in mind. The only solution is that those of us working in ESP/EAP use the potential offered by corpus linguistics to build our own dictionaries. The experimental *Parasitic Plant Dictionary, PPD,* is an attempt to do just this.

6. Building meanings

29 The information provided by corpora led to a need for a reappraisal of dictionary content, the ordering and presentation of senses, with their grammatical and lexical context. As Rundell (1998) has pointed out the advent of large electronic corpora changed our outlook on source materials, not only in terms of the quantity of data, but also the quality of the data. The sheer size of modern reference corpora has lead to the development of new tools designed to assist in the handling of such large quantities of data. However, tools do not replace lexicographers and their knowledge of language, quite the contrary, they provide a means to control the data leaving the lexicographer with the time to apply his or her expertise in the elaboration of more meaningful and precise entries (Rundell 2002).

30 Although computational lexicography has become an area in which complex routines abound (see Ooi 1998 for background information), the basic tool used in all corpus-based lexicography remains the concordancer. However, it is all too easy to look at concordance lines and forget the philosophical approach on which such a tool relies.

31 One important notion is that of representativity. Although this is attained by size and careful selection in reference corpora, the question is much more difficult in EAP/ESP corpora (Williams 1999, 2002b). Any corpus project must take into account sociolinguistic criteria; reliance on mere statistical selection would merely render it a mass of data. Whilst representativity remains a contentious issue, we must at least be able to justify our content, and thereby the conclusions drawn from that content. Bearing this aspect in mind we must consider what the concordance lines are actually showing. According to Tognini Bonelli (2001) the concordance lines from a correctly constructed corpus overcome the Saussurean *langue/parole* divide. The individual lines are clearly syntagmatic, representing *la parole*, individual instances of uses, whilst the paradigmatic whole gives us *la langue*, the collective knowledge of language representative of a living language community. Thus, what we endeavour to capture in a dictionary is no longer

some speculative notion of *sense* and *meaning*, but something firmly anchored in the reality of usage. In such a paradigm *meaning* must be realised in context.

32 In corpus lexicography we must accept that *sense* and *meaning* must be clearly differentiated. *Sense* covers a variety of notions, including the encyclopaedic and the semantic. *Sense* is essentially a conceptual notion and much of what is essential in an encyclopaedic entry will not be found in a corpus as it relies on wider world knowledge. However, the linguistic means by which these senses are communicated can be approached in a corpus and grouped in dictionary entries. These will help the user to decode, to understand, but not necessarily to encode. Semantic analysis can help in encoding, but is applying a human categorisation which may not be readily recognisable by the user. The semantic approach can be seen in WordNet (<www.cogsci.princeton.edu/~wn/>) in which words are analysed within a hierarchy. The senses are clearly delimitated, but a learner may be none the wiser. It is through example that dictionaries such as the COBUILD endeavour to leap the divide between *sense* and *meaning.*

33 Whilst *sense* is an abstract notion, *meaning* can only be approached through context. This is the Firthian approach to language which underpins all learner's dictionaries, meaning can only be seen through instantiation, and each instantiation is unique. Hence the conundrum that every dictionary must seek to resolve, if meaning is contextual, then a dictionary is not. How can we realise this transfer of meaning from instantiated context into a decontextualised word book and then back into new instantiated meaning? The answer has been shown for the COBUILD, and now all learner's dictionaries, but what about in specialised usage.

34 If we accept that concordance lines can overcome the langue/parole divide, and if we accept that concordances can help us find senses through analysis of meaning, then we have the tools to build our own dictionaries.

35 There are many concordancers on the market, some free, some commercial, but none expensive. That produced by Mike Scott, *WordSmith Tools*, is probably the most widely used by lexicographers working within a Windows environment. Now coming available as version four, *WordSmith Tools*, is much more than just a concordancer, offering a wide variety of options for lexical analysis including the capacity to handle morphosyntactic (Part of Speech -POS) analysis and Text Encoding Initiative (TEI) compliant corpora. In this example we shall only look at concordancing and collocational profiles.

36 The first stage is obviously to build a corpus of texts. The design criteria are paramount as an *ad hoc* corpus can only lead to *ad hoc* results. Whilst dictionaries based on large reference corpora can claim representativity, the keyword in specialised corpora is justification, being clear as to what has gone into the corpus and why (Williams 2002b). The next stage is the selection of headwords. This depends largely on the perceived purpose of the dictionary, if the aim is to show terms in context then a manual or automatic system of term extraction will be needed, in the case of the *PPD* a wider lexical base is sought using collocational networks (Williams 2001) as a means to reduce subjectivity and to avoid an over-reliance on simple frequency, which would exclude most specialised items. Whatever the method adopted the most important step is the use of the concordancer to extract the individual senses of the word. This will be demonstrated through the example of "*control*".

7. Getting into Control

37 In Williams 2002, I argue for the necessity of specialised learner's dictionaries to allow users to visualise senses in an environment close to their own working environment as the generalised senses found in dictionaries do not provide examples that will easily enable these users to instantiate *their* meanings in *their* working context. The problem with terminologies is that they essentially address the translator, not the subject specialist, specialists "know" their terminology and do not necessarily wish to follow prescriptive advice, but to create new contextualised meanings. Consequently if we attempt to enter the field with a prescriptive approach, or even a descriptive one, we are entering a potential mine field, the answer is to build a dictionary to show contextualised usage, specialised *parole*.

38 The problem can be illustrated through the word *control.* This semi-technical word is not covered by any of the senses given in the COBUILD online and whilst its main technical use is signalled by *(techn)* in OALD 5, the definition is addressed to the general user. If we look at a semantically organised database as WordNet (http://cogsci.princeton.edu/~wn/), we find 11 senses of which two, senses 1 and 4 could clearly be relevant to a technical context. However Wordnet provides no syntactic, nor collocational information. Coming back to our corpus, the first step in looking for meaning in context will be to build a concordance for this node word.

39 Quite apart from left and right sorting which will reveal syntagmatic units, the obvious next step in sense disambiguation is to sort by part of speech. If meaning is linked to syntax then verbs and nouns will provide differing sense patterns. Once this basic division has been carried out, more precise analyses can be carried out. For the noun the next stage might be looking at singular and plural forms separately as these can generate very different meanings, for example, in the case of control, the sense of "*used as for comparison*" is predominantly singular. This sorting out of senses is a gradual process wherein syntactic choices help define meaning choice and *vice versa*. This is corpus-driven lexicography (Williams 2002a) in which the analysis is driven by the content and emerging patterns, not the intuitions of the compiler. As the sense patterns are built up we must also look for restricted collocations as there is no point in building an encoding dictionary if these be absent. WordSmith provides collocational profiles, which are a precious aid in pattern building and word sense disambiguation by presenting the cooccurence patterns around a node as a sortable table. Phrase patterns can also easily be seen using the cluster facility.

40 Working through the definition for *control* from the *PPD* (appendix 1), we can see clearly how the sense divisions are made. First comes the verb and noun forms. The collocational profile then helps break the verb form into the two main senses found in this specialised context. The examples show the patterns that have emerged, in sense 1. X *controls* Y by Z, the preposition is an important key in locating this sense, when it is not present X still *controls* Y. In the second sense the *control* is exercised over a process, *growth* in the example. The present participle *controlling* obviously has a sense that is related to sense 1, so is given here. The pattern is different, but again the sense differentiation is helped by the presence of a preposition. The same process is followed for the noun forms, each sense being shown with its phrasal patterns and collocations. These patterns are essential to word sense disambiguation, and in turn are essential in the dictionary so as to assist in

the encoding process. No staggering new senses are revealed here, such senses have already been isolated by standard lexicographical intuition, that is not the point. What has been achieved is to show the senses that really occur in a given environment, by frequency of use and within the type context that someone working in this field would understand, and need to reproduce. The definitions given are minimal, this is deliberate. The aim is to show how the words may be used, not to get bogged down into prescriptive definitions. The standardising of terms is useful, but the negotiation of meaning is done through contexts.

41 The ordering of the entry is made to take into account the electronic format. The majority of dictionaries are designed to be printed, the online version or cd-rom versions tend to follow the print layout. Print dictionaries are expensive to produce and cannot evolve rapidly without new editions, so the logical choice for an ESP/EAP learner's dictionary in rapidly changing field is a web-based dictionary using hyperlinks. Hyperlinking means that a pre-entry can be used to show the main subdivisions of the entry. Within each sense we get a short definition, followed by examples of use, collocations and phrasal patterns. Entries may be more or less long, the advantage of electronic dictionaries is that the constraints imposed by a paper presentation no longer hold, which means that much finer detail may be obtained. The process is a lengthy one, but stimulating. It is obvious that such fine detail cannot be achieved in commercial dictionaries where time is of the essence; however, in the world of ESP/EAP we need not have such considerations in mind. However, even for unfunded specialised corpora more advanced techniques of analysis will gradually speed up the task.

8. Going Further

42 As Rundell (2002) has pointed out, computers will never replace man, but in practical lexicography they can help considerably. The stage forward in meaning extraction does not only mean improving tools, but also the corpora themselves. There is obviously a need for clearly thought out selection criteria, both internal and external (Williams 2002b), and we shall certainly benefit from better corpus annotation.

43 Part-of-speech (POS) tagging is now a fairly standard process, albeit relatively time-consuming. The main problem with any morphosyntactic analysis is that of error due to lack of training; the tools must be "taught" to work in a specific textual environment. Once training has been done, analysis is fast, and can only be sped up as more sophisticated tools become available. However, whilst time spent on analysis can be justified in the building of reference corpora, which have a long shelf life, the same may not be true for specialised corpora which, to follow the evolution of science, require regular updating. There is another danger in that in POS tagging, one distances oneself from the text by imposing rules that may hinder analysis. As all linguists know, parts of speech are purely artefacts and do not correspond to real entities that are valid in all cases, especially as concerns prepositions and conjunctions. Nevertheless POS tagging does offer numerous advantages, amongst which are the identification of syntactic patterns that would otherwise be hidden by the mass of data. In turn POS tagging combined with text markup can lessen the arduous task of the lexicographer in seeking lexical patterns and reducing the ambiguity induced by polysemy.

44 WordSketch (Kilgarriff & Rundell 2002) is part of a lexicographer's workbench that has been designed to run on the British National Corpus, that is to say a fully POS tagged, TEI

compliant corpus. Thus far the technology is within the reach of any corpus linguist, the next stage goes further in partially parsing the corpus.

9. WordSketch

45 Part of speech annotation simply names the words, but does not tell us what syntactic role they are playing. It is now fully accepted that meaning and syntax are totally intertwined (Sinclair 2002), which must mean that in analysing lexis in context we need to see syntactic patterns. Parsing seeks to add this information to a corpus.

46 Full parsing is fraught with difficulty given the extreme complexity of real data as opposed to the cleanliness of a grammarian's model. It also means accepting a grammatical model, and one model does not fit all. The answer in the WordSketch project is to partially parse by setting out to annotate a series of patterns on annotated and lemmatised corpus. The initial project outlined 26 patterns into which a keyword may enter (Kilgarriff & Tugwell 2001). To see what a Wordsketch does we can take the example of *control*.

47 Wordsketches are built online (see http://wasps.itri.bton.ac.uk for a demonstration). The tool offers a number of search parameters, the two that concern us here are search word and part of speech. With control as search word we find a choice between three options; noun, verb, adjective. For each part of speech a sketch is built showing the relevant patterns. For example, *control* as noun (appendix 2.) is associated with prepositional phrases as *PP_of_situation*, or *PP_over_ situation*. *Situation* is one of a list of words associated with this particular pattern. Similarly, *control* may be a modifier, as in *remote, tight* or *strict,* or be modified as in *control samples*. Verb collocations are shown as *object_of*. For each pattern the associated words are shown with their frequency, clicking on the words brings up a concordance. Working through the patterns allows us to isolate the meanings by associating the lexical and syntactic environments in relation to other patterns, for instance the pattern *noun_modifier_pest* links with *PP_obj_method_of* to give *methods of pest control*. The same process can be seen with the verb patterns which reveal relationships such as prepositional phrase, subject, object or modifier. *Modifier* will give the adverbial collocations such as strictly *controlled*. It must be borne in mind that the corpus has been lemmatised so all forms of the verb or noun can be represented.

48 This is a very sophisticated system built using a reference corpus and as such is beyond the means of the EASP/EAP corpus builder. However, corpus linguistics is about comparison, so cross-checking a variety of sources is the key to understanding. The results of a *Wordsketch* can be checked out on the BNC itself to see exactly which genre make use of this formula, the meanings can also be confirmed by looking at *WordNet*. Patterns found in a specialised corpus can be cross-checked to a *Wordsketch*, and vice verse.

49 There are, however, dangers in over-reliance on technology. Computers cannot replace the lexicographer, but an inexperienced researcher can easily be blinded by science. Wordsketch, and other word sense disambiguation projects, are designed to speed up the work of a lexicographer, but speed can also mean that interesting material is overlooked. Using such tools means to accept the validity not only of the part of speech markup, but also a grammatical analysis. The result is a corpus-based analysis which essentially confirms findings, in research this must be associated with a corpus-driven approach

building up patterns and meanings by critical observation. No one approach is perfect, all have their advantages and drawbacks, the linguist must simply be aware of these.

Conclusion: Practising what one preaches

50 This text has set out to show the possibilities of corpus lexicography in building specialised learner dictionaries using the example of the *Parastic Plant Dictionary*. Why, it might be asked, does the experimental parasitic plant dictionary not exist as more than a few pages on a web site? The question needs to be asked as I may seem to not be practising what I preach, the answer is of course time, and academic status.

51 To begin with the latter. I was told many years ago when setting out upon a thesis that a dictionary was not a thesis. This is true, but the result in many cases is that we are obliged to talk about things that we do not actually do in practice. Dictionary criticism can be the topic of a thesis and the basis of an academic career: writing dictionaries is not. This of course leads to the question of time as any young researcher, no matter their physical age, gets caught into an academic treadmill where only theoretical papers count. If we want to improve language teaching in ESP/EAP through specialised lexicography this situation must come to an end, real questions need real, not theoretical answers and this will only come about by trial and error. It is obvious that the creation of specialised learner's dictionaries will never really interest the major dictionary houses as they can only handle large marketable resources; the answer must lie elsewhere. In ESP/EAP teaching, major changes have come about by a mixture of non-commercial theoretical and practical research, the cooperation amongst teachers on the ground. Surely the same thing can be done in lexicography. Metaphor may be an interesting subject of a thesis, but however exciting the cognitive model we adopt it will not give rapid results in the ground, changes in lexicographical practice could. This is not to write off metaphor or any other theoretical study, but to ask that practical research be given equal status.

52 To conclude I turn to the father of modern lexicography, Dr Samuel Johnson with his celebrated definition of a lexicographer:

> Lexicographer: a writer of dictionaries, a harmless drudge.

53 Practical lexicography is a highly time-consuming pastime, but given the easy access to computers and the rise of corpus studies in ESP/EAP research there is no reason why it could not be developed. Drudgery it is, sexy it is not, but the value of the results in teaching terms could be immeasurable.

I wish to express my thanks to all members of the lexicographical community, notably members of the EURALEX board who have helped through encouragement and criticism. In particular I wish to thank Adam Kilgarriff and David Tugwell of ITRI, University of Brighton, for giving me access to the full WordSketch database. The careful proof reading of the anonymous reviewers has led to the clarification of some points and the removal of some idiocies. I thank them for their patience.

BIBLIOGRAPHY

Websites

COBUILD Online: http://www.linguistics.ruhr-uni-bochum.de/ccsd/

COBUILD: http://titania.cobuild.collins.co.uk/

EURALEX: http://www.ims.uni-stuttgart.de/euralex/

WordNet: http://www.cogsci.princeton.edu/~wn/

WordSketches: http://wasps.itri.bton.ac.uk

WordSmith Tools: http://www.liv.ac.uk/~ms2928/

Parasitic Plant Research Dictionary: Please note that this is an experimental dictionary and therefore incomplete. Not all the information is online and updates are irregular.

Dictionaries

OALD : *Oxford Advanced Learner's Dictionary*

COD : *Concise Oxford Dictionary*

ODBMB : *Oxford Dictionary of Biochemistry and Molecular Biology*. 2000. Smith, A.D., S.P. Datta, G. Howard Smith and P.N. Campbell. New York: Oxford University Press.

RCS : *Roberts and Collins Senior*

Other Works

Atkins, B.T.S, J. Clear, N. Ostler. 1992. "Corpus design criteria". *Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing* 7/1, 1-16

Atkins, B.T.S. 2002. *Bilingual Dictionaries: past, present, future*. In Corréard, M-H. (ed.), 1-29.

Béjoint, H. 2000 [1994]. *Modern Lexicography: An introduction*. Oxford: Oxford University Press.

Biber D. 1993. "Representativeness in corpus design". *Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing* 8/4, 243-257.

Braasch, A., C. Povlsen (eds.) 2002. *Proceedings of the 10th EURALEX International Congress. August 13 - 17. 2002.* Denmark: Center for Sprogteknologi.

Campoy Cubillo, M.C. 2002. "Dictionaries use and dictionary needs of ESP students: An experimental approach". *International Journal of Lexicography* 15/3, 205-228.

Correard, M.-H. (ed.). 2002. *Lexicography and Natural Language Processing: A festschrift in honour of B.T.S. Atkins*. EURALEX <http://www.euralex.org/>.

Cowie, A.P. 1998. "A.S. Hornby. 1898-1998: A centenary tribute". *International Journal of Lexicography* 11/4, 251-268.

Daille B. and G. WILLIAMS (eds.). 2001. *Proceedings of the Collocations Workshop*. Toulouse: ACL.

Fontenelle, T. 2002. "Lexical knowledge and Natural Language Processing". In Corréard, M.-H. (ed.), 216-229.

Howatt A.P.R. 1984. *A History of English Language Teaching*. Oxford: Oxford University Press.

Kilgarriff, A., and M. Rundell. 2001. "Wordsketch: Extraction and display of significant collocations for lexicography". In Daille B. and G. Williams (eds.), 32-38.

Kilgarriff, A., and M. Rundell. 2002. "Lexical profiling software and its lexicographical applications - A case study". In Braasch A. and C. Povlsen (eds.), 859-864.

McCreary, D.R.. 2002. "American freshmen and English dictionaries: I had aspersions of becoming an English teacher". *International Journal of Lexicography* 15/3, 181-205.

McCreary, D.R. and F.T. Dalezal. 1999. "A study in dictionary use by ESL students in an American University". *International Journal of Lexicography* 12/2, 99-108.

Norman, G. 2002. "Description and prescription in dictionaries of scientific terms". *International Journal of Lexicography* 15/4, 259-276.

Ooi, V.B.Y. 1998. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Quirk, R, S.Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rundell, M. 1998. "Recent trends in English pedagogical lexicography". *International Journal of Lexicography* 11/4, 315-342.

Rundell, M. 2002. "Good old fashioned lexicography: Human judgement and the limits of automation". In Corréard (ed.), 138-155.

Tognini Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamin's Publishing

Sinclair, J. (ed.). 1987. *Looking Up: An account of the COBUILD Project in Lexical Computing*. London: Collins.

Sinclair, J. 2002. "Lexical Grammar". *Tekstynu Lingvistika, Darbai ir Dienos* 24, 191-203.

Williams, G. 1999. *Looking in before looking out: Internal selection criteria in a corpus of plant biology.* Papers in Computational Lexicography. Complex '99. Hungary : Budapest. : 195-204.

Williams, G. 2001. *Mediating between lexis and text: collocational networks in specialised corpora. ASp* 31-33, 63-76.

Williams, G. 2002a. "Corpus-driven lexicography and the specialised dictionary: headword extraction for the Parasitic Plant Research Dictionary". In Braasch A. and C. Povlsen (eds.), 859-864.

Williams, G. 2002b. "In search of representativity in specialised corpora: categorisation through collocation". *International Journal of Corpus Linguistics* 7/1, 43-64.

APPENDIXES

Appendix 1. An example entry: *Control*

control

This word may be used as a verb or a noun. The forms controlling and controlled can serve as adjectives.
verb , *transitive. to control. control, controls, controlling, controlled*
The verb *control* has two main uses, to limit or restrain and to manage

Sense 1. To control something is to restrain or limit the capacity to act of something. In the case of parasitic plant biology it means to limit the spread or degree of infestation of the parasites.

examples

1. Most farmers in Gambia control *Striga* by weeding
2. The tree canopy completely controlled *Striga* infestation
3. *Orobanche aegyptiaca* was controlled by all chemigated treatments

The verb is often modified by a modal verb such as *can* or *may*
• Pot experiments had shown that chlorosulfuron can control Broomrape in tomato
• Ethylene may control the growth of dodder

Phrasal patterns. The present participle *controlling* is often used in the pattern of adjective/ noun + in + controlling + noun or noun + for + controlling + noun
• The herbicide was found effective in controlling the parasite
• The semi-arid zone of West Africa holds a great potential in controlling*Striga hermonthica*

• they could provide a potential for controlling parasites

Sense 2. to control is to manage something
• Ethylene negatively controls the growth of dodder

The noun *control* has three main meanings: as a a comparative sample in an experiment,a means to restrain something and to manage something.

Sense 1, *noun, countable*. **control** as comparative sample. In an experiment, an uncontaminated sample or population is used as a standard against which the infected sample can be compared.
examples
1. A set of uncovered plants was used as a control
2. In control experiments, RNA probes were subjected to the RNAse protection protocols

Collocation. The noun *control* is more frequently found in the singular, where it is refered to a *a control* (general) or *the control* (specific). The singular form may be used as a modifier as in *control plants* or be modified as in *the untreated control*. The plural form cannot modify another noun, but can be modified as in *uninfected controls*.
Control can modify nouns.
• control plants - *infected plants used less water than control plants.*
• control maize - *infected maize plants were significantly shorter than control maize.*
• control tissue - *Control tissues consisted of non-inoculated roots.*
• control lanes - *Control lanes show that there is no template activity.*

Control and *controls* can be modified by adjectives.
• negative control - *The fungal Pesta served as negative control*
• positive control - *Orobanche seeds served as positive control*
• susceptible control - *Cultivar Peredovik was used as a susceptible control*
• untreated control - *Seeds were similar to the untreated control*
• uninfected controls - *the combined dry weights were similar to that of uninfected controls*

Phrasal patterns. Control is often used in comparative phrases. Different patterns require different prepositions.
• *Height was lower than that of control plants.*

• *The number of Striga plants was significantly lower than in the control*
• *dry mass was slightly modified in comparison with the control*
With the verb *compare, to* or *with* may be used.
• *A higher number of Striga plants compared to the control*

• *Each treatment was compared with the control*

Sense 2, *noun, generally singular*.control as restraint or limitation. *Control* in this sense is the action to prevent the spreading or propagation of parasite rather than their elimination. *Control* concerns parasitic plants as weeds rather than botanical specimens . examples
1. Satisfactory control can be achieved with glyphosate.
2. Hand weeding is still the best control treatment.
3. All that is required is to keep the weeds in control is prevention of seeding.

Collocation. The noun is often modified by the name of parasitic weed, for example*Striga control*, or by the word *weed*. The noun is frequently found modifying another noun to form a term.
Control as modifier
• control agent(s) - *The potential of natural enemies as biological control agents has recieved much attention.*
• control approaches- *traditional control approaches have been inadequate.*
• control cells - *control cells had been bombarded with plasmids*
• control measure(s) - *Several control measures have been employed.*
• control method(s) - *Several control measures have been proposed.*
• control packages - *The break-even incremental yield of this control package must be 43 to 55kg per hectare.*

Control can be modified by adjectives expressing adequacy.
• good control - *glyphosate has shown good control of Broomrape*
• better control - *knowledge of the taxonomy and biology of parasites should lead to better control*

• best control - *imazapyr provided the best control*
• complete control - *The herbicide gave complete control of Broomrape*
• excellent control - *moderate to excellent control of Orobanche was achieved*

Control can be modified by adjectives expressing suitability of the process.
• effective control - *crop rotation could be an effective control method*
• possible control - *farmers should be trained in possible control methods*

Control can be modified by adjectives expressing a variety or number of approaches.
• *different* control methods were used
• *several* control runs were performed
• *various* control strategies have been developed

Control can be modified by adjectives expressing type of approach.
• biological control - *The same fungus was evaluated for biological control of Striga hermonthica.*

• chemical control - *A new experimental approach to the chemical control of Striga.*
• post/pre-emergence control - *An effort to identify effective chemicals for the post-emergence control of the parasite.*

Control is often associated with certain verbs
• achieve - *almost complete control was achieved*
• demonstrate - *Brown (1991) demonstrated control of Orobanche*
• establish - *Host plant resistance is likely to be the most successful means to establish control of the parasite*

Sense 3, *noun, generally singular*.control as management *Control* in this sense is the action to manage the presence of something. It is often expressed the phrase *to be under the control of something or someone . Control* concerns parasitic plants as weeds rather than botanical specimens .

Collocation. The noun may be modified by an adjective and is frequently found modifying another noun to form a term. This sense is closely related to the sense of limiting something.

Control modified by an adjective
• integrated control - *An integrated control approach is needed*
• sustainable control - *The development of effective and sustainable control measures*
• long-term control - *an integrated long-term control approach is needed*

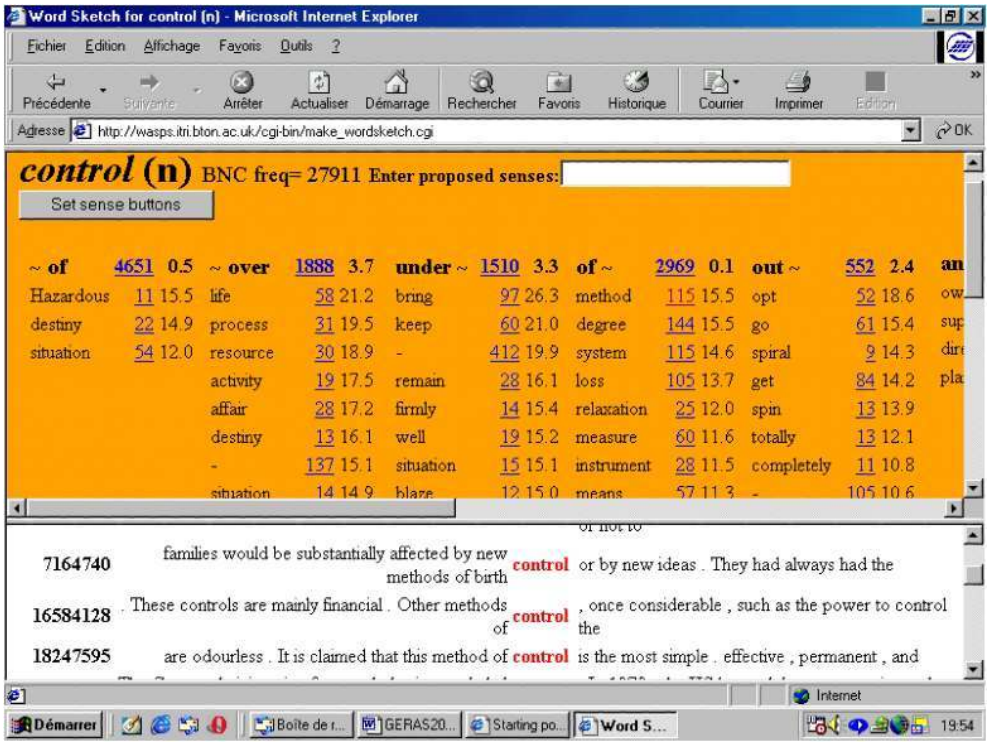
Control modifying a noun
• control authorities - *Australian state or local weed control authorities eradicate or contain other species.*
• control program(s)/programme(s) - *This herbivore could have some potential in biological control programmes.*
• control strategy/strategies - *Various chemical control strategies have been developed in the USA.*
• control technology/technologies - *Past research efforts developed a diversity of control technologies.*

Phrasal patterns.
• these movements are under the control of endogenous rhythms
• Synthesis may *be under* phytochrome *control* in higher plants

controlling, adjective. *Controlling* refers to something that controls.
• This could be the controlling factor

controlled, adjective. *Controlled* refers to something that is controlled.
• The experiments were carried out under controlled conditions

Appendix 2. A partial WordSketch for Control



ABSTRACTS

Corpus linguistics has revolutionised lexicography leading to better learner's dictionaries. Dictionary senses are decontextualised senses, but learner's dictionaries have been evolved to help with both decoding and encoding. However, general language dictionaries do not necessarily meet the needs of ESP users as the transfer of sense from a general to specific context is difficult. This text shows how monolingual learner's dictionaries have evolved and how language corpora have influenced them. The article discusses the problems of poor dictionary skills and shows how lexicographers attempt to overcome this through clearer word sense disambiguation. The writer shows how senses may be extracted from specialised corpora with the aim of building a specialised ESP encoding dictionary.

La linguistique de corpus a révolutionné la lexicographie et a conduit à de meilleurs dictionnaires d'apprentissage. Les sens dictionnairiques sont hors contexte, mais les dictionnaires d'apprentissage ont évolué pour aider au décodage et à l'encodage. Cependant, les dictionnaires de langues généraux ne sont pas nécessairement adaptés aux utilisateurs LANSAD puisque le transfert d'un sens général à un contexte précis est difficile. L'article démontre comment les dictionnaires d'apprentissage ont évolué sous l'influence de la linguistique de corpus. Il traite du problème du manque de compétences dans l'utilisation des dictionnaires et démontre la manière dont les lexicographes ont essayé de surmonter les difficultés d'utilisation en adoptant de meilleures méthodes de désambiguïsation du sens. L'auteur présente des techniques pour l'extraction du sens d'un corpus pour la création d'un dictionnaire d'encodage spécialisé.

INDEX

Keywords: corpus linguistics, learner's dictionary, lexicography
Mots-clés: dictionnaire d'apprentissage, lexicographie, linguistique de corpus

AUTHOR

GEOFFREY WILLIAMS

Geoffrey Williams est maître de conférences à l'Université de Bretagne Sud, Lorient où il enseigne l'anglais en LEA. Il enseigne également la linguistique de corpus en licence et DEA des Sciences du langage à l'Université de Nantes. Il est membre de nombreuses associations internationales en linguistique de corpus et en lexicographie. Il fait partie du CRELLIC, le Centre de Recherche en Littératures, Linguistique et Civilisations à Lorient et de la jeune équipe ALPL, Analyse Linguistique et Pratiques Langagières à Nantes. geoffrey.williams@univ-ubs.fr